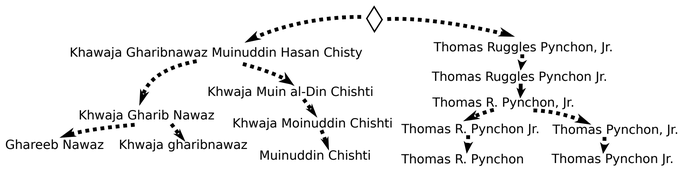
Name Phylogeny: A Generative Model of String Variation

Nicholas Andrews, Jason Eisner, Mark Dredze

Department of Computer Science, CLSP, HLTCOE, Johns Hopkins University

A Example Name Phylogeny



Edges from the root correspond to generating new name strings (entities)

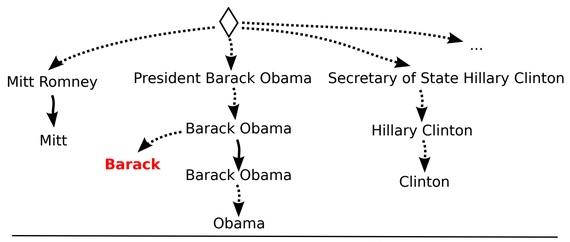
Edges between strings are "mutations"

Generative model

Given a sequence of tokens, there are two options to generate the next name mention (token):

1. Pick an existing token x with probability 1 / (α+k)
   1.1 **Copy** x verbatim with probability 1 - μ
   1.2 **Mutate** x with probability μ
2. **Generate** a new string with probability α / (α+k)

Generative model in action



x_{10001} = Mitt Romney
x_{10002} = President Barack Obama
x_{10003} = Barack Obama
x_{10004} = Secretary of State Hillary Clinton
x_{10005} = Hillary Clinton
x_{10006} = Barack Obama
x_{10007} = Clinton
x_{10008} = Obama
x_{10009} = Mitt
x_{10010} = Barack

Mutation model

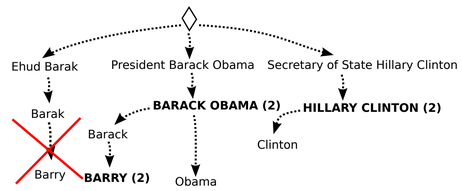
We use a simple conditional PFST with latent edit and no-edit regions.

The basic character-level operations are insertions, deletions, substitions, and copies.

The parameters θ of the model determine the probabilities of transitions between latent regions and of the different character-level operations.

From token phylogenies to type phylogenies

All **copy** edges are collapsed (see vertices in bold below)



The first token in each collapsed vertex is a mutation; the rest are copies

Approximation: we forbid multiple tokens of the same type to be derived from mutations

Inference via EM

Iterate until convergence:

1. **E-step:** Given θ, compute a distribution over phylogenies (spanning trees)
2. **M-step:** Re-estimate transducer parameters θ given marginal edge probabilities

Inner EM loop in the M-step to sum over alignments between input/output strings

Data

Wikipedia redirects are used as ground truth for entity name variations

The frequency of each name variant is estimated using the **Google/Stanford crosswiki dataset** (Spitkovsky and Chang, 2012)

For evaluation, 500 entities are sampled and their name **tokens** are divided into **5 training** folds and **1 test** fold

The training dataset contains ~4000 distinct strings

Supervision constrains the phylogeny

Subsets of the training folds restrict possible spanning trees:

(1) Edges only allowed between labeled types of the same entity and unsupervised types.
(2) No edges from unsupervised types to supervised types.

(Both supervised and unsupervised types may derive from the root.)

Experiments

At training time:

1. Estimate transducer parameters with EM
2. Find the single best phylogeny given the learned model

At test time:

1. Attach test tokens to the inferred phylogeny
2. Calculate precision and recall for the connected component the test token was attached to

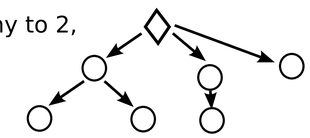
Precision: fraction of name variants in the connected component of the same entity as the test token

Recall: fraction of all name variants for the test token found in the connected component

Baseline: Flat Tree

As a baseline, we limit the depth of the phylogeny to 2, so each name variant either

(1) mutates from a fixed canonical name or
(2) is generated from scratch (the root)



The most frequent name variant for each entity is selected as the canonical name.

Results

For the **baseline** and the **full model**, we vary:
- the proportion labeled data at training time
- the parameter α